Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Refinement of protein crystal structures using energy restraints derived from linear-scaling quantum mechanics

Ning Yu,[a] Hemant P. Yennawar[b] and Kenneth M. Merz Jr[a]*

[a]Department of Chemistry, The Pennsylvania State University, 104 Chemistry Research Building, University Park, Pennsylvania 16802, USA, and [b]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 8A Althouse Laboratory, University Park, Pennsylvania 16802, USA

Correspondence e-mail: merz@psu.edu

Received 28 October 2004
Accepted 17 December 2004

A novel method is proposed in which combined restraints derived from linear-scaling semiempirical quantum-mechanical (QM) calculations and X-ray diffraction data are combined to refine crystal structures of proteins. Its performance has been tested on a small protein molecule, bovine pancreatic trypsin inhibitor (BPTI). The refinement involves minimization of the sum of a geometric energy function and an X-ray target function based on either the least-squares residual or the maximum-likelihood formalism. For comparison, similar refinement runs have also been performed using energy restraints derived from the force field available in the *Crystallography & NMR System* (*CNS*) program. The QM refinements were carried out with weights that were varied by several orders of magnitude and the optimal weights were identified by observing the trend in the final free *R* values, QM heats of formation and coordinate root-mean-square deviations (r.m.s.d.s) from the crystal structure. It is found that the QM weights are typically smaller but generally on the same scale as the molecular-mechanics (MM) weights for the respective X-ray target functions. The crystallographic *R*, free *R*, real-space *R* values and correlation coefficients based on the structures refined with the energy restraints derived from our QM calculations and Engh and Huber parameters are comparable, suggesting that the QM restraints are capable of maintaining reasonable stereochemistry to a similar degree as the force-field parameters. A detailed inspection of the structures refined with the QM and MM energy restraints reveals that one of the common differences between them and the crystal structure is that the strained bond angles in the crystal structure are corrected after energetically restrained refinements. Systematic differences in certain bond lengths between the QM-refined structures and the statistical averages of experimental structures have also been observed and discussed.

1. Introduction

Three-dimensional structural information has been shown to be vital in understanding protein function and elucidating reaction mechanisms. To date, the vast majority of protein structures have been determined by single-crystal X-ray crystallography. However, despite many technological advances such as synchrotron-radiation sources and low-temperature techniques on the experimental side and simulated-annealing (SA) refinement and maximum-likelihood formalism (MLF) on the theoretical side, true atomic level resolution protein structures achieved with X-ray diffraction are still very rare (Wlodawer *et al.*, 1984; Jelsch *et al.*, 2000). This is to a great extent a consequence of the limited

diffraction data with respect to the number of structural parameters required to model coordinates and thermal effects of atoms in protein crystals. In the X-ray experimental community, this problem has been traditionally dealt with by introducing constraints or restraints (Jack & Levitt, 1978; Hendrickson, 1985; Tronrud *et al.*, 1987) during refinement. The purpose of the former is to reduce the number of adjustable parameters, whereas the latter essentially increases the amount of observations by supplementing the X-ray data with stereochemical information. Generally, the lower the resolution of the X-ray data, the more heavily the restraints are weighted in the refinement. The typical restraints are very similar to the energy terms in molecular-mechanics force fields, which are energy terms representing bond lengths, bond angles, chirality, planarity and nonbonded repulsion (Hendrickson, 1985; Brunger & Adams, 2002) and are weighted in such a way that the deviations match those found in databases of high-resolution structures. The main difference between these energy restraints and standard force fields for proteins such as CHARMm and AMBER is that electrostatic interactions and van der Waals attractions are not retained (Brunger & Adams, 2002) mainly because of the difficulty in their parameterization, although efforts have been made recently to reintroduce electrostatics into protein X-ray structure refinement (Moulinier *et al.*, 2003). In the commonly used macromolecular structure-solution program *Crystallography & NMR System* (*CNS*; Brünger *et al.*, 1998), the parameters of the energy restraints are derived from a statistical analysis of the chemical moieties of proteins and polynucleotides from the Cambridge Structural Database (CSD) by Engh & Huber (1991). These parameters have been shown to be very accurate in refining the structures of protein molecules. However, when the system under investigation contains a ligand, inaccurately determined parameters for the ligand can lead to significant structural errors in the final model. Therefore, it is desirable to have restraints that are less dependent on atom typing and parameterization, but are still able to represent the physical interactions.

Quantum-mechanical methods have been used to study the electronic structures, equilibrium geometries and thermodynamic properties of molecules both in the gas phase and in the condensed phase, usually with a continuum description of the solvation environment. Currently, the available computational power has limited the applications of most *ab initio* and density functional theory (DFT) methods to small molecules of up to a couple of hundred atoms. Nevertheless, there has been continuous development in electronic structure methods whose computational costs scale linearly with system sizes (Yang, 1991; Yang & Lee, 1995; Lee *et al.*, 1996; Dixon & Merz, 1996, 1997). These methods have extended the upper bound of the SCF type of theory to calculations of full protein systems of a few thousand atoms. For example, Alsenoy *et al.* (1998) have optimized the geometry of the protein crambin *in vacuo* at the HF/4-21G level of theory; Liu *et al.* (2001) have employed a linear-scaling semiempirical molecular-orbital (MO) method to carry out molecular-dynamics simulations of the same protein molecule with a quantum-mechanical treatment of the entire protein and a molecular-mechanical description of solute–solvent interactions.

Using quantum-chemistry calculations, the total charge density or the difference density can be studied. In experimental X-ray structure refinement, the principal model that is used to represent the electron density for a structure during refinement is the independent atom model (IAM), which assumes (i) the electron-density function $\rho(x, y, z)$ is a superposition of the densities of the atoms that comprise the unit cell and (ii) the atomic densities are described by the spherically averaged densities of isolated ground-state atoms. In actuality, charge-density distributions in real-world molecules are different from the superpositions of independent atomic densities as a result of chemical bonding between atoms. Thus, the deviation from the IAM, namely the deformation density or difference density, is an interesting subject that can be studied by quantum-chemistry calculations. As shown by many theoretical studies (Chandler *et al.*, 1994, 2000; Catti & Pavese, 1996; Krijn *et al.*, 1988; Krijn & Feil, 1988; Swaminathan *et al.*, 1984) and by us recently in an SCF study of the charge-density distributions of the glycyl-L-threonine dihydrate molecule (Yu & Merz, 2004), deformation densities usually account for a difference of less than 1% in R values even at ultrahigh resolutions. Thus, the use of the IAM as the phasing model should not be the dominant source of error for the relatively high R values obtained for protein structures compared with those in small molecules. Evidently, the reason often has more to do with the problems that can ultimately be attributed to the sizes of macromolecules: *e.g.* proteins are dynamic molecules but current refinement models can only to a very limited degree account for their flexibility, phasing is considerably more difficult and may be biased by the starting structures and protein refinements often involve a large number of model parameters, which can cause significant optimization problems. Despite these difficulties, deformation-density analyses on protein crystals have been performed on crambin (Jelsch *et al.*, 2000) and scorpion toxin (Housset *et al.*, 2000), which have demonstrated the existence of the valence-density deformation features also observed in charge-density studies on small molecules.

Recently, the application of quantum-mechanical calculations to validate protein models in the crystalline state has attracted much attention. Ryde and coworkers first proposed a method, *ComQum-X*, that combines quantum-mechanical/molecular-mechanical (QM/MM) calculations with crystallographic raw data to refine the structures of substrates in the presence of protein environments (Ryde *et al.*, 2002; Ryde & Nilsson, 2003*a*,*b*,*c*). In this approach, the ligands were treated by quantum-mechanical calculations using DFT and the proteins and solvent were modeled by the force-field parameters in *CNS* and AMBER. Since the force field alone may not be sufficient to maintain the structure to be compatible with the crystal data, an X-ray penalty function is introduced as a pseudo-energy term in the total target function. In one of these investigations studying the crystal structure of oxidized *Bacillus pasteurii* cytochrome c_{553} (Ryde & Nilsson, 2003*b*), these authors reported success in refining the lower resolution

structure to one that was closer to an available higher resolution structure and the R value of the *ComQum-X*-refined structure was lowered by 0.6 to 1.1% based on the reflections observed for the lower resolution structure. It has to be noted, however, that this scheme is different from conventional QM/MM methods (Warshel & Levitt, 1976; Field *et al.*, 1988) in that since the electrostatics of the MM region is not treated, the QM–MM interaction Hamiltonian contains only repulsive Lennard–Jones potentials. Thus, the structure of the ligand at the binding site is determined by the collective forces arising from the QM/MM energy and the X-ray diffraction data, the latter of which are actually related to the structure of the whole binding complex.

In this paper, we propose a method that is similar to the one that *ComQum-X* implements to refine the crystal structure of a pure protein molecule, bovine pancreatic trypsin inhibitor (BPTI). BPTI is a small 58-residue protein molecule whose structure has been studied extensively using a number of experimental techniques such as X-ray diffraction (Wlodawer *et al.*, 1984; Deisenhofer & Steigemann, 1975; Huber *et al.*, 1970), neutron diffraction (Wlodawer *et al.*, 1984) and nuclear magnetic resonance (Wuthrich *et al.*, 1982). The major difference that distinguishes our approach from *ComQum-X* is that the whole protein is treated quantum mechanically using our linear-scaling semiempirical (SE) electronic structure program *DivCon*, which we will describe briefly in the next section. Since the structure of the entire BPTI molecule is refined, important structural changes emerge as a result of energy relaxation, which are accompanied by changes in crystallographic R values and real-space correlation coefficients. These changes and final structures will be compared to the crystal structure as well as the structures refined with the force field in *CNS*. Our main objective in this study is to study the competence of semiempirical quantum-chemical calculations as a refinement tool. Thus, we wish to see whether the energy restraints derived from full SE-QM calculations, when combined with X-ray data, can yield protein structures with comparable R values and stereochemical quality as those refined with conventional MM restraints in *CNS*. Furthermore, we would like to identify systematic trends in the stereochemical parameters that can be expected from structures after QM-restrained refinements.

2. Methods

In crystal structure refinement involving the use of energy restraints, an energy function based on the chemical information is combined with an X-ray target function to be minimized (Jack & Levitt, 1978),

$$E_{\text{total}} = E_{\text{chem}} + w_{\text{X-ray}} E_{\text{X-ray}}, \tag{1}$$

where E_{total} is the function to be minimized, E_{chem} is the geometric energy function, $E_{\text{X-ray}}$ is the X-ray target function and $w_{\text{X-ray}}$ is the weight, which we will discuss more later. In the commonly used refinement force fields such as that in *CNS*, E_{chem} is a sum of the terms describing various types of interactions, *i.e.*

$$\begin{aligned} E_{\text{chem}} = {} & \sum_{\text{bonds}} k_b (b - b_0)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 \\ & + \sum_{\text{dihedrals}} k_\varphi \cos(n\varphi + d) + \sum_{\text{chiral,planar}} k_\omega (\omega - \omega_0)^2 \\ & + \sum_{i<j} a r_{ij}^{12}, \end{aligned} \tag{2}$$

where the familiar attractive van der Waals and electrostatic terms have been omitted. In contrast, in the QM refinement scheme, E_{chem} is the quantum-mechanical energy E_{qm}, or more precisely the QM heat of formation, obtained from MO calculations,

$$E_{\text{total}} = E_{\text{qm}} + w_{\text{X-ray}} E_{\text{X-ray}}, \tag{3}$$

which inherently takes into account all the nuclear–electronic interactions. In this work, E_{qm} is evaluated on the semiempirical AM1 level using our linear-scaling divide-and-conquer algorithm.

The straightforward choice for the X-ray target function is the least-squares residual,

$$E_{\text{X-ray}} = E^{\text{LSQ}} = \sum_{hkl} w[|F_{\text{o}}(hkl)| - k|F_{\text{c}}(hkl)|]^2. \tag{4}$$

However, as pointed out by Read and others (Brunger & Adams, 2002; Read, 1986, 1990; Pannu & Read, 1996; Adams *et al.*, 1997), use of the residual as the target function is only justified for models that are very close to the true structure, which is often not the case in macromolecular refinement. An improved target function can be derived using the maximum-likelihood formalism (MLF),

$$E_{\text{X-ray}} = E^{\text{MLF}} = \sum_{hkl} \left(\frac{1}{\sigma_{\text{ML}}^2} \right) [|F_{\text{o}}(hkl)| - \langle |F_{\text{o}}(hkl)| \rangle]^2, \tag{5}$$

which is more suitable for the general case of incomplete models and models that contain initial bias. In this work, we have employed both residual and MLF target functions and compared their performance. During the refinement process, the progress towards convergence is monitored by the residual index, or the R value, defined as

$$R = \frac{\sum_{hkl} w \big| |F_{\text{o}}(hkl)| - k|F_{\text{c}}(hkl)| \big|}{\sum_{hkl} w |F_{\text{o}}(hkl)|}, \tag{6}$$

where w is a weighing factor associated with each reflection, $|F_{\text{o}}|$ is the experimentally observed structure factor and $|F_{\text{c}}|$ is that calculated from the model. An improvement from traditional refinement practices, which use all the observed reflection data to refine structures, is a procedure that includes cross-validation, as suggested by Brünger (1992). In this procedure, about 5–10% of the reflections are set aside for the purpose of validating the model and excluded from the working set of X-ray diffraction data used in minimization. The R value computed for this reserved set of reflections is called the free R value and is an unbiased indicator of how well the structural model fits the experimental observations. Since $E_{\text{X-ray}}$ in (1) is a dimensionless quantity, the weight at which the X-ray diffraction data is factored into the total target function is an arbitrary parameter. Although automatic

procedures exist to determine a sensible weight in a typical refinement program such as *CNS*, ultimately the only objective criterion for choosing the weight is to choose the one that minimizes the free R value.

2.1. The divide-and-conquer (D&C) strategy

Our scheme to utilize energy restraints derived from QM calculations is enabled by a linear-scaling electronic structure technique, the divide-and-conquer method. For more details on this method, the reader is referred to previous theoretical work (Dixon & Merz, 1996, 1997; van der Vaart *et al.*, 2000). Here, we will only give a brief outline of its theory and strategy.

The typical implementation of the Hartree–Fock (HF) self-consistent field (SCF) formalism involves an iterative solution of the Roothaan–Hall equation (Hehre *et al.*, 1986; Szabo & Ostlund, 1996),

$$\mathbf{FC} = \mathbf{SCE}, \tag{7}$$

where **F** is the Fock matrix, **C** is the MO coefficient matrix, **S** is the overlap matrix and **E** is the diagonal eigenvalue matrix. In the framework of the semiempirical MO theory (Dewar, 1969), **S** reduces to the identity matrix as a consequence of the neglect of diatomic differential overlap (NDDO) approximation. This gives rise to the simplified semiempirical Roothaan–Hall equation,

$$\mathbf{FC} = \mathbf{CE}, \tag{8}$$

which is solved to variationally minimize the electronic energy with respect to the MO coefficients. The quantity that is central to our divide-and-conquer scheme is the density matrix **P** of the whole molecular system, defined as

$$P_{\mu\nu} = 2\sum_{i=1}^{N_{\rm occ}} c_{\mu i}c_{\nu i}, \tag{9}$$

where $c_{\mu i}$ and $c_{\nu i}$ are the expansion coefficients of MO i in the basis of atomic orbitals χ_μ and χ_ν and the summation is over all the doubly occupied MOs for closed-shell species. Once **P** is determined, the total energy is computed by summing the electronic and nuclear energies,

$$\begin{aligned} E_{\rm total} &= E_{\rm elec} + E_{\rm nuc} \\ &= \frac{1}{2}\sum_{\mu=1}^{N_{\rm elec}}\sum_{\nu=1}^{N_{\rm elec}}(\mathbf{H}_{\mu\nu} + \mathbf{F}_{\mu\nu})\mathbf{P}_{\mu\nu} + \sum_{\alpha}^{N_{\rm atom}}\sum_{\beta<\alpha}\frac{Z_\alpha Z_\beta}{R_{\alpha\beta}}, \end{aligned} \tag{10}$$

where **H** is the one-electron core Hamiltonian (Szabo & Ostlund, 1996; Levine, 2001) whose elements are defined as

$$\mathbf{H}_{\mu\nu} = \int \chi_\mu^*(\mathbf{r})\left(-\frac{1}{2}\nabla_{\mathbf{r}}^2 - \sum_{\alpha}^{N_{\rm atom}}\frac{Z_\alpha}{|\mathbf{r}-\mathbf{R}_\alpha|}\right)\chi_\nu(\mathbf{r})\,{\rm d}\mathbf{r} \tag{11}$$

and **F** is the same Fock matrix as shown in (7) and (8) and is defined as

$$F_{\mu\nu} = H_{\mu\nu} + \sum_{\lambda=1}^{N_{\rm elec}}\sum_{\delta=1}^{N_{\rm elec}}[(\mu\nu|\lambda\delta) - \tfrac{1}{2}(\mu\delta|\lambda\nu)]P_{\mu\nu}, \tag{12}$$

where the following notation has been used for the two-electron integrals,

$$(\mu\nu|\lambda\delta) = \int\int \chi_\mu^*(\mathbf{r}_1)\chi_\nu(\mathbf{r}_1)\frac{1}{|\mathbf{r}_1-\mathbf{r}_2|}\chi_\lambda^*(\mathbf{r}_2)\chi_\delta(\mathbf{r}_2)\,{\rm d}\mathbf{r}_1 d\mathbf{r}_2. \tag{13}$$

The tremendous speeding up in the NDDO-type methods is mainly attributed two reasons: (i) the one- and two-electron integrals are approximated by semiempirical expressions using atom-based parameters as opposed to being evaluated explicitly and (ii) many of the electron-repulsion integrals (ERI) vanish owing to the NDDO approximation, including all the three- and four-center integrals, which constitute the majority of the ERIs. Since $E_{\rm tot}$ depends on **F** and **P**, and **F** also depends on **P**, while **P** in turn depends on the MO coefficients, which are the solution of (7) or (8), the SCF process is an iterative one in which an initial guess $\mathbf{P}^{(1)}$ of the density matrix is formed first, which is used to assemble the first Fock matrix $\mathbf{F}^{(1)}$, which is then diagonalized to yield the first set of MOs. The coefficients of these MOs are used to form an improved density matrix $\mathbf{P}^{(2)}$, which is again used to generate the next Fock matrix $\mathbf{F}^{(2)}$. These procedures are repeated until the maximum change in the elements of the density matrix between two successive steps is smaller than a certain threshold value, when SCF reaches convergence.

The key point in the divide-and-conquer strategy, similar to many other sparse-matrix techniques, is to divide the whole system into a set of subsystems, each with their own core and buffer regions. This is realised by dividing the global density matrix into smaller density matrices of subsystems and solving the Roothaan–Hall equation locally in each of the subsystems,

$$\mathbf{F}^\alpha\mathbf{C}^\alpha = \mathbf{C}^\alpha\mathbf{E}^\alpha, \tag{14}$$

which yields a density matrix $\mathbf{P}^\alpha$ for each of the subsystems,

$$P_{\mu\nu}^\alpha = \sum_{i=1}^{N^\alpha}\frac{2}{1+\exp[(\varepsilon_i^\alpha - \varepsilon_{\rm F})/kT]}c_{\mu i}^\alpha c_{\nu i}^\alpha, \tag{15}$$

where α labels the subsystem, N^α is the total number of MOs in subsystem α, $\varepsilon_{\rm F}$ is the Fermi energy and ε_i^α is the eigenvalue of MO i in that subsystem. In our implementation, the subsystem core regions do not overlap one another and the buffer regions consist of two layers, each of which can overlap any part of another subsystem. The global density matrix is now assembled from subsystem density matrices as

$$P_{\mu\nu} = \sum_{\alpha=1}^{N_{\rm sub}} D_{\mu\nu}^\alpha P_{\mu\nu}^\alpha, \tag{16}$$

where

$$D_{\mu\nu}^\alpha = \begin{cases} 1/n_{\mu\nu} & \text{if } \chi_\mu \text{ is in the core and } \chi_\nu \text{ in the core or inner buffer, or } \textit{vice versa} \\ 0 & \text{otherwise} \end{cases}, \tag{17}$$

where $n_{\mu\nu}$ is the total number of systems that overlap to make a nonzero contribution to the global density-matrix element $P_{\mu\nu}$. Therefore, the SCF convergence in a D&C calculation requires two conditions: (i) in every subsystem, the Roothaan–Hall equation must be solved to yield converged subsystem density matrices and (ii) the global density matrix, after being

assembled from subsystem contributions, must show convergence between two successive steps.

2.2. Computational details

Our refinement scheme is implemented by a Perl script that drives both *DivCon* and *CNS* and alternates between the two programs. *CNS* is a general-purpose structure-solution program that has a highly flexible hierarchical architecture, *i.e.* it provides a symbolic structure-determination language (*CNS* language) and a set of task-oriented input files. We have scripted the conjugate-gradient minimization file that came with the *CNS* distribution, minimize.inp, in such a way that it does not perform minimization but only evaluates and outputs the X-ray target function and gradient based on the input structure. The motivation for this is because *CNS* only treats nonaliphatic H atoms and thus we decided to use *DivCon*, which treats all atoms explicitly, to merge the energy and X-ray gradients and to update the coordinates. This task fits naturally into a geometry-optimization process, which is one of the many functions that *DivCon* provides. The assumption underlying this practice is that the X-ray gradient changes more slowly than that of the energy, which has been justified by earlier work on energy refinement (EREF; Jack & Levitt, 1978) and refinement by simulated annealing (Brünger *et al.*, 1989). Before the calculations start, a script is run first to construct a mapping table that relates the atoms in *CNS* to those in *DivCon*. The files that are required by our script are all the necessary inputs to *DivCon* and *CNS* and several additional files that pass the gradients and coordinates between the two programs. The script itself consists of a loop that iteratively calls the two programs and checks the outputs for convergence. Our criteria for convergence are that the difference between the QM energies at two successive steps must be less than 0.4 kJ mol^{-1} and the all-atom coordinate r.m.s.d. must be less than 0.001 Å. Moreover, the maximum component of the total gradient in the last step has to be smaller than a certain threshold value. Based on extensive computational experiments, we found that 20.9 kJ mol^{-1} Å^{-1} is most reasonable for refinement runs with relatively low weights on the X-ray data and it was adjusted accordingly to up to 41.9 kJ mol^{-1} Å^{-1} for runs with larger $w_{\text{X-ray}}$.

The initial structure of BPTI was taken from the Protein Data Bank (PDB code 5pti) and was processed so that the QM refinement could be carried out. Firstly, the original PDB structure contains an unknown ion at site 324, which was hypothesized to be potassium. It also has a phosphate ion with an occupancy of 0.65 and 63 water molecules, 34 of which are partially occupied. These heteroatoms are modeled crystallographically as part of the molecular model; however, the partial occupancies make them difficult to treat quantum mechanically at this time. Furthermore, since the focus of the present study is to compare the energy restraints derived from QM calculations with those based on the Engh and Huber parameters for proteins, the inclusion of the phosphate ion and water molecules, for which additional parameters would have to be used, would complicate the comparison. Therefore, we decided to follow the approach adopted in the simulated-annealing (SA) refinement studies (Brünger *et al.*, 1987, 1989) and strip off the phosphate ion and water molecules from the model. Although this treatment results in a slight increase in the *R* and free *R* values, it is expected to have relatively small effects on the refined protein structures, as shown in the results obtained by Brünger *et al.* (1987). The presence of alternate conformations for Glu7 and Met52 creates another complication for our calculations and for simplicity in this work we have decided to retain only the side-chain conformations with the largest occupancies and assign them occupation numbers of 1.0, as performed in other theoretical investigations such as in Ryde *et al.* (2002). Additionally, the original PDB file contains some D atoms, for which specific semiempirical QM parameters do not exist (in cases where D is important, for example vibrational frequencies, D is incorporated *via* mass weighting only). In the present study, these atoms have been replaced by H atoms because (i) the X-ray scattering factor for hydrogen is equal to that for deuterium since they are isoelectronic and (ii) the changes in bond lengths between D atoms and neighboring heavy atoms are expected to be minimal and beyond the resolution at which the diffraction data were collected.

The X-ray data were also downloaded from the PDB and include 17 615 reflections between the resolution limits of 1.0 and 8.0 Å. The experimental paper reports an *R* value of 0.200 based on a model with individual anisotropic temperature factors, which are not available from the PDB, and the *R* value computed for a model with the equivalent isotropic temperature factors (*B* factors) is 0.2078 with bulk-solvent correction (or 0.2093 without). Since the original data set does not label the reflections used in cross-validation, we decided to randomly select 892 (5%) reflections to form our own free *R* set. We also decided not to perform bulk-solvent correction in the subsequent calculations. This decision is motivated by two considerations: (i) again, we wish to assess the differences between the QM energy restraints and the MM ones and reduce as much as possible the other factors that might interfere with the comparison and (ii) since the lower resolution limit of the current data set is 8.0 Å, bulk-solvent correction should not cause a significant difference. Thus, for our starting structure, which only contains part of the crystallographic model without any ions or water molecules, the *R* and free *R* values are 0.2571 and 0.2700, respectively. These *R* values were based on the agreement between experimental structure factors and model structure factors computed with the Fast Fourier Transform (FFT) method, which might introduce an extra source of error, especially at higher resolutions. However, as shown by comparison between the results of *CNS* refinement and those from our own code employing the atomic scattering-factor formalism to compute the structure factors, the differences were quite small and did not change the results significantly. Finally, in the present work the individual isotropic temperature factors were not refined but held fixed at their values in the PDB file, as was performed by Brünger *et al.* (1987, 1989). Refining the individual *B* factors does not create any theoretical obstacles, but may obscure the

Table 1
The R values, free R values, quantum-mechanical heats of formation and all-atom r.m.s.d.s from the crystal structure for QM refinement runs with varying weights and the two X-ray target functions.

	MLF target function					Residual target function			
Weight	R	R_{free}	E_{qm} (kJ mol^{-1})	R.m.s.d. (Å)	Weight	R	R_{free}	E_{qm} (kJ mol^{-1})	R.m.s.d. (Å)
1.0×10^{0}	0.2528	0.2799	−5494.1	0.187	1.0×10^{5}	0.2543	0.2957	−4323.9	0.217
5.0×10^{-1}	0.2545	0.2781	−5962.5	0.198	5.0×10^{4}	0.2572	0.2924	−5600.2	0.241
4.0×10^{-1}	0.2553	0.2779	−6128.5	0.236	4.0×10^{4}	0.2571	0.2926	−5874.0	0.250
3.0×10^{-1}	0.2564	0.2785	−5240.6	0.247	3.0×10^{4}	0.2586	0.2907	−6101.1	0.260
2.0×10^{-1}	0.2580	0.2775	−6279.0	0.200	2.0×10^{4}	0.2601	0.2898	−6284.5	0.238
1.0×10^{-1}	0.2641	0.2779	−6431.2	0.201	1.0×10^{4}	0.2648	0.2911	−6550.0	0.269
1.0×10^{-2}	0.3025	0.3264	−6908.6	0.308	1.0×10^{3}	0.3000	0.3259	−6882.4	0.299

differences between the QM- and MM-refined structures incurred by the different energy restraints, since the effects resulting from the changes in atomic coordinates can usually be compensated by adjusting B factors.

3. Results

Starting from the initial structure, we refined the structure of BPTI with quantum-mechanical energy restraints. For comparison, we also carried out refinement runs using the molecular-mechanical energy restraints available in *CNS*. In MM refinements, *CNS* has a built-in procedure to determine the optimal weight of the X-ray target function automatically. This is not the case for QM refinements, where we had to vary the weight by several orders of magnitude and use the free R value as a guide to choose the most suitable weight. For validation, we use the R and free R values to evaluate the overall protein structures and the local correlation R values and electron-density maps to study the detailed structural changes resulting from different refinement protocols.

3.1. Determination of the optimal weight ($w_{X\text{-}ray}$)

In a typical *CNS* refinement run, a short molecular-dynamics (MD) simulation is performed first to obtain an estimate of the energy and X-ray gradients and then $w_{X\text{-}ray}$ is chosen so that the average magnitudes of the two gradients are approximately equal. This is not possible in QM refinement, which is computationally more expensive than force-field methods. Thus, our approach is to perform QM refinements for different weights and make the choice based on the final free R values of the refined structures. The optimal weights determined by the automatic procedure in *CNS* were 0.208952 and 29759.2 for the MLF and residual targets, respectively, which provided us with an initial estimate. For the MLF target, the weight was varied between 1 and 0.01, whereas the range for the residual target was from 1000 to 100 000. In Fig. 1 we plot the quantum-mechanical heat of formation as a function of the R value during the refinements when the MLF target was used. The plots for other refinements are qualitatively similar and thus are omitted from presentation. With both targets, the heat of formation of the protein molecule was lowered by 6697–7535 kJ mol^{-1}. The effect of the X-ray target on the R values is noticeable: as shown in Figs. 1(*a*) and 1(*b*), for refinement runs with the heaviest weights $w_{X\text{-}ray}$, the X-ray target is lowered simultaneously with the QM energy. However, subsequent examination of the corresponding free R values, which are collected in Table 1, clearly indicates over-fitting of the X-ray data in these runs. Table 1 also shows the final R values, QM energies and coordinate root-mean-square displacement (r.m.s.d.) of all the atoms of the refined structure from the starting structure. As $w_{X\text{-}ray}$ is gradually lowered, the final R values show moderate increase and yet the free R values become lower, while the structures are further relaxed in terms of the final energies. For the residual target, this trend is clearly discernable and the optimal weight for the QM refinement was estimated to be 20 000 (the fifth series) with a final free R value of 0.2898. For the MLF target, despite the fact that the final free R value of the refinement run with $w_{X\text{-}ray} = 0.3$ is abnormally higher than the neighboring ones, our criterion suggests that 0.2 is the optimal weight. These choices are consistent with the trend observed for the r.m.s.d.s., *i.e.* the r.m.s.d.s for the heaviest weights are the lowest for the runs with the respective X-ray targets and start growing when the weights are lowered. However, when the weights reach the optimal values the coordinate r.m.s.d.s fall into their local minima. Similarly, we performed refinement runs with energy restraints derived from the force field parameterized by Engh and Huber using the *CNS* program. We varied the weights in the same range as

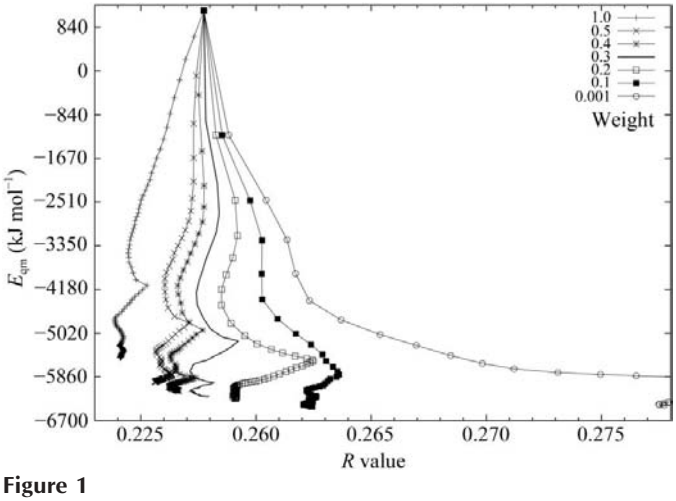

Figure 1
The quantum-mechanical heat of formation against the R value for the QM refinement using the maximum-likelihood formalism (MLF) X-ray target with weight varied between 0.01 and 1.0.

Table 2
The R and free R values for the crystal, QM-refined and *CNS*-refined structures.

Residual index	Crystal	QM/MLF (wa = 0.2)	QM/residual (wa = 20000)	*CNS*/MLF (wa = 0.8)	*CNS*/residual (wa = 40000)
R (over 16 723 reflections)	0.2571	0.2580	0.2601	0.2593	0.2616
Free R (over 892 reflections)	0.2700	0.2775	0.2898	0.2761	0.2961

used in the QM refinements and the optimal weights were found to be 0.8 for the *CNS*/MLF combination and 40 000 for *CNS*/residual combination. It is interesting to note that in the QM/MM refinement of *N*-methylmesoporphyrin-bound ferrochelatase by Ryde *et al.* (2002), the authors suggested that the *CNS* MM forces are typically three times larger than their QM forces, which seems consistent with our observations.

3.2. Overall protein structures

Since the solvent and alternate conformations were omitted during the refinement and the atomic B factors were not refined as part of the model, the structures after QM and MM refinements have cosmetically higher R values than the crystal structure. However, as discussed in §2.2, the main focus of this work is to assess the performances of different energy restraints in assisting refinement. The working R and free R values based on the structures after refinement with QM and MM energy restraints using maximum-likelihood formalism (MLF) or residual X-ray target functions are displayed in Table 2 and we also bin these R and free R values into ten ranges of resolution and show the plot in Fig. 2, where the corresponding values for the crystal structure are included only for reference purposes. It can be seen that at resolutions lower than 1.7 Å ($\sin\theta/\lambda < 0.3$), the R and free R values for the MLF target are very close, whereas the R values for the residual target are consistently larger but are close between QM and MM refinements, the only exception being the R values between 2.14 and 8 Å ($0.06 < \sin\theta/\lambda < 0.23$). This is partly because the lower resolution reflections are more sensitive to the structure on the large scale, *i.e.* the fold, secondary structure *etc.*, while the energy restraints mainly control the stereochemical details. In this case, the X-ray target function is more important than the energy restraints in maintaining the agreement with X-ray data. This result is clearly a reflection that the MLF target is superior to the residual one, especially considering that between 2.14 and 8 Å the R values for the MLF target are higher but the free R values are actually lower. This trend continues for the R values as the resolution increases to up to 1 Å ($\sin\theta/\lambda = 0.5$), whereas for the free R values this is less obvious. This can be explained in a number of ways: (i) the reflection regions in the higher resolution shells contain fewer reflections than lower resolution shells owing to data incompleteness, which can give rise to poorer statistics, and (ii) the higher angle reflections contain larger experimental errors. The increases in the final R values of various refined structures from that of the crystal structure are in the third decimal place, but the increases in the free R values are considerably larger, between 0.0061 and 0.0261. The free R value of the QM/MLF structure is higher than the *CNS*/MLF structure by 0.0014, whereas the same energy function with the residual target gave a free R value that was better by 0.0063.

The lowering in energy of the QM refinements was mainly attributed to reorientations of the H atoms to form more favorable interactions and considerable adjustments of main-chain and side-chain stereochemical parameters to relieve strain. To further investigate the plausibility of the structures, we used the *PROCHECK* program (Laskowski *et al.*, 1993) to obtain the key indicators for the stereochemical quality and collect the results in Table 3. Apparently, the QM refinements improve the percentage of residues in the core region of the Ramachandran plot over the *CNS* refinements. From the G factors, the QM-refined structures show relatively large deviations of the φ–ψ, χ_1–χ_2, χ_1, χ_3, χ_4 and ω dihedral angles from the statistical distribution derived from 163 protein

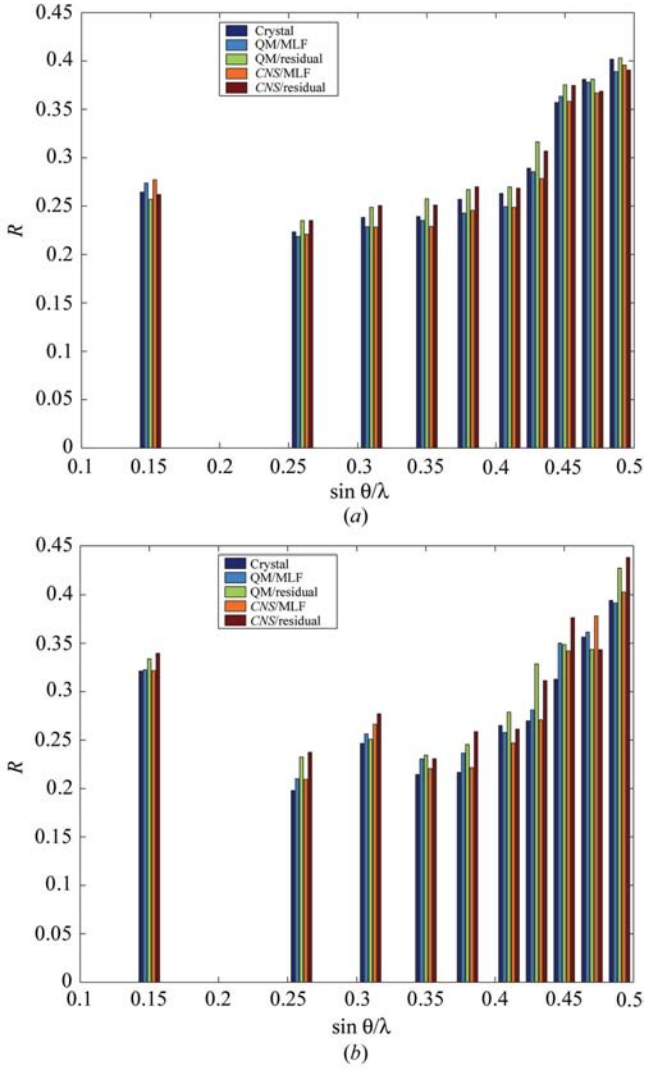

Figure 2
The R (*a*) and free R (*b*) values as functions of the scattering vector $\sin\theta/\lambda$ for the working set of 16 723 reflections (*a*) and the free set of 892 reflections (*b*).

Table 3
The stereochemical quality indicators obtained with the *PROCHECK* program for the crystal, QM-refined and *CNS*-refined structures.

Indicators	Crystal	QM/ MLF	QM/ residual	*CNS*/ MLF	*CNS*/ residual
Ramachandran (%)					
Core	91.30	93.50	95.70	91.30	89.10
Allowed	8.70	6.50	4.30	8.70	10.90
G factors					
Dihedrals	0.08	−0.33	−0.34	0.12	0.13
Covalent	−1.71	−0.43	−0.50	0.42	0.42
Overall	−0.58	−0.39	−0.50	0.25	0.25

structures solved by X-ray crystallography to a resolution of 2.0 Å or better and an *R* factor of no larger than 20%. Nevertheless, the plausibility of the main-chain bond lengths and bond angles, shown by the covalent *G* factors, of the QM structures was improved relative to the crystal structure, although the improvement was not as considerable as that from the MM refinements.

3.3. Local structural changes

As we have demonstrated, the MLF target function has a clear advantage over the residual one for refinements with both the QM and MM energy restraints. Therefore, in the subsequent analysis on local structural changes, we will only focus on the structures refined with the MLF target function.

Firstly, we examined the QM-refined structures further to locate regions in which the QM refinement led to systematic differences from the ideal values in the MM parameters of Engh and Huber. It was discovered that the CZ—NE, CZ—NH1 and CZ—NH2 bond lengths of all the Arg residues are longer than the ideal values by 0.052–0.081 Å, the CB–SG bond lengths of all the Cys residues are shorter than the ideal values by 0.051–0.077 Å and the Asp3—Phe4, Pro9—Tyr10 and Gly56—Gly57 C—N peptide bond lengths are longer than the ideal values by 0.054–0.060 Å. In fact, the parameters for NE—CZ bonds are the same as those for C—N peptide bonds in *CNS*, with an equilibrium bond length of 1.329 Å and a force constant of 12 643 kJ mol^{-1}Å^{-2}, and these bonds are consistently longer in the QM-refined structure by 0.052–0.062 Å. The energy penalties associated with these bond-length deviations are relatively low, between 17 and 50 kJ mol^{-1} as estimated with the parameters in *CNS*. Interestingly, the corresponding ideal values in the AMBER force field for these bond lengths are very similar to the Engh and Huber parameters, which would suggest that this is a systematic problem of the AM1 method owing to its parameterization scheme. However, correcting this deficiency would require a re-parameterization of the semiempirical Hamiltonian and is not the focus of the present work.

In Fig. 3 we plot the real-space *R* values (Fig. 3*a*) and real-space correlation coefficients (Fig. 3*b*) residue by residue for the QM/MLF structure and the *CNS*/MLF structure, where the highly disordered Gly57 and Ala58 have been omitted. For comparison, we have also included the curve for the crystal structure. We see that, since the temperature factors are not refined and the solvent was not included, the real-space *R* values for the crystal structure are lower than those for the other two structures. The most problematic regions include the surface residues Arg1, Asp3, Lys15, Arg17, Lys26, Lys41, Glu49, Asp50 and Arg53, which show the same trend for all three structures, indicating that the systematic differences in the bond lengths observed in QM refinement do not have a significantly negative effect on the local electron-density correlation. In particular, all Cys residues (5, 14, 30, 38 and 51) of the QM/MLF structure have equal or better real-space *R* values than those of the *CNS*/MLF structure and the density around Cys30 fits the observations even better than that of the crystal structure, despite the fact that the CB—SG bond lengths of all the Cys residues in the QM structure are shorter than the statistical average.

A few bond angles in the crystal structure deviate from the ideal values by a considerable amount and are corrected by energy restraints in both QM and MM refinements, as shown

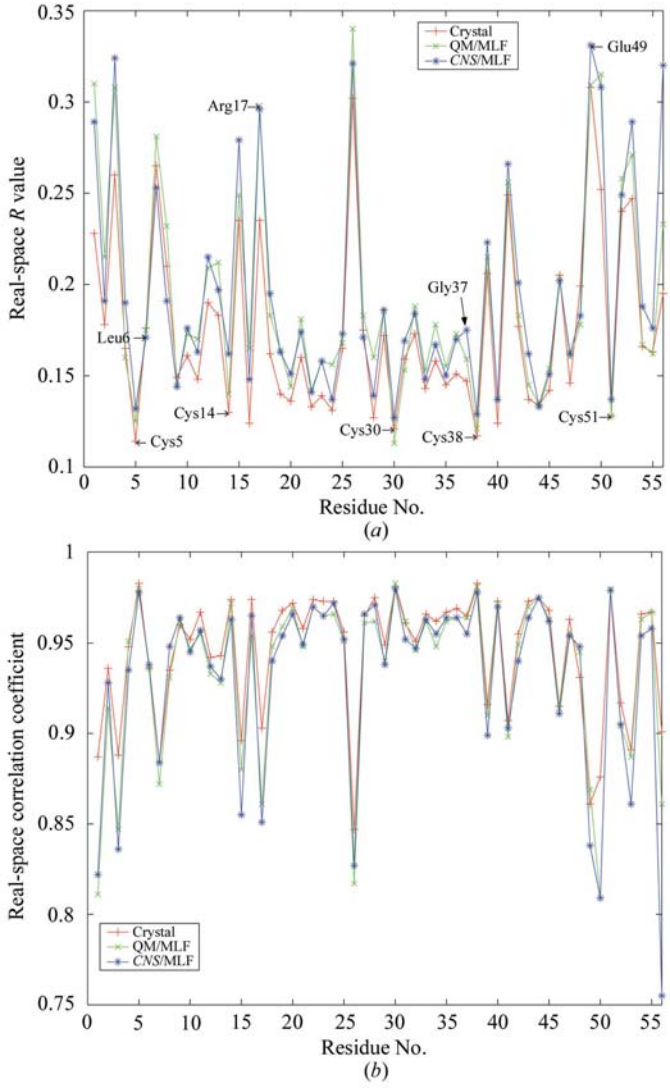

Figure 3
The real-space *R* values and correlation coefficients *versus* residue numbers for the crystal, QM/MLF and *CNS*/MLF structures.

Table 4
Selected bond angles in the crystal, QM/MLF and *CNS*/MLF structures and the Engh and Huber ideal values for θ_0 and K_θ.

Bond angles		Crystal	QM/MLF	*CNS*/MLF	θ_0	K_θ
Arg17	CD—NE—CZ	153.2	125.1	124.7	124.2	863.744
Glu49	CB—CG—CD	130.7	112.0	113.3	112.6	672.465
	CG—CD—OE1	139.2	120.1	118.8	118.4	367.377
	OE1—CD—OE2	94.8	120.8	122.7	122.9	337.400
Leu6	CA—C—O	111.7	119.7	120.4	120.8	672.465
Gly37	N—CA—C	121.6	116.5	118.5	112.5	231.085

in Table 4. An example of this correction is the CD—NE—CZ angle of Arg17, which is shown in Fig. 4 together with the σ_A-weighted $2F_o - F_c$ electron-density maps based on the crystal and the QM/MLF structures. In the crystal structure, this angle is 153.2°, almost 30° greater than the equilibrium value of 124.2°, which causes an energy penalty of 929 kJ mol^{-1} evaluated with Engh and Huber parameters. As shown in Fig. 4, the corresponding electron-density map computed for the crystal structure seems to justify the placement of the NE atom, which appears inside the mesh contoured at the 1.1σ level. After restrained refinement, the NZ atom is displaced so as to relax the CD—NE—CZ angle, which is about 125° in both the QM/MLF and *CNS*/MLF structures. The map computed for the QM/MLF structure is not significantly different from that for the crystal structure, except that there is no density for the NZ atom in the crystal map. This suggests that in disordered regions such as Arg17, the phases used in the construction of the electron-density map are strongly influenced by the model structure, so much so that alternative interpretations of the observed data can arise. In order to ascertain the most reasonable structure, we further computed the simulated-annealing omit map and present the result in Fig. 4. In calculating the omit map, we used the 'complete' molecular model including all the alternate conformations, the phosphate ion and the solvent molecules and only omitted the atoms that are within a radius of 3.5 Å from the center of the residue in question. The omit map for most of Arg17 is similar to those for the crystal and QM/MLF structure, but the main difference appears in the region around the guanidine group. It can be seen from Fig. 4 that the omit map does not cover the NE atom in the crystal structure, providing an argument against the strained CD—NE—CZ angle. From Fig. 3, the restrained refinement results in an increase of the local R value by about 0.6, which could be partially compensated if the temperature factors of the atoms in the guanidine group were refined accordingly. However, this was not undertaken for the reasons explained in §2.2. No major difference is observed between the QM/MLF and *CNS*/MLF structures and they give very similar local R values and correlation coefficients.

In Fig. 5, we plot the structures and electron-density maps of Glu49 in a similar way to Fig. 4. In the crystal structure, the

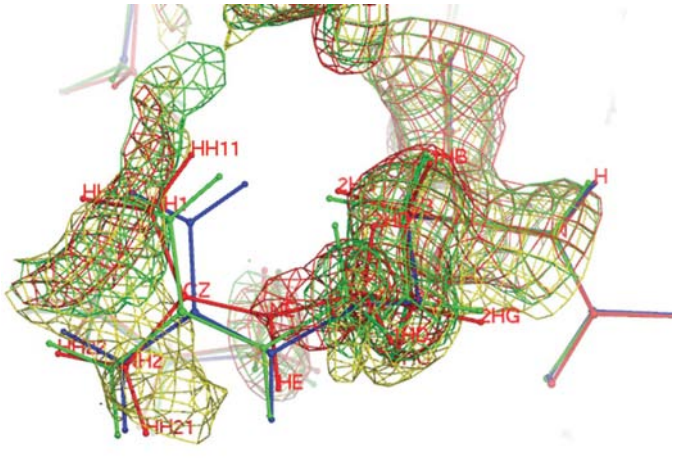

Figure 4
The Arg17 residue of the crystal structure (red), the QM/MLF structure (green) and the *CNS*/MLF structure (blue), together with the σ_A-weighted $2F_o - F_c$ electron-density maps based on the crystal structure (red), the QM/MLF structure (green) and the omit map (yellow). All maps are contoured at the 1.1σ level.

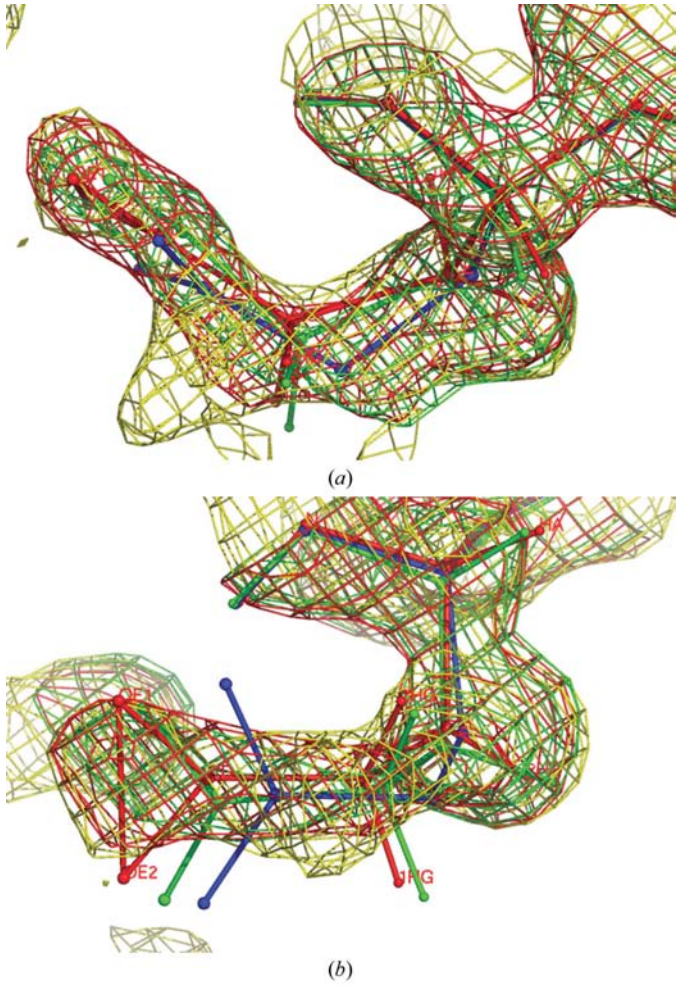

Figure 5
The Glu49 residue of the crystal structure (red), the QM/MLF structure (green) and the *CNS*/MLF structure (blue), together with the σ_A-weighted $2F_o - F_c$ electron-density maps based on the crystal structure (red), the QM/MLF structure (green) and the omit map (yellow). All maps are contoured at the 1.1σ level. In (*a*) the CD—OE1—OE2 plane is almost perpendicular to the plane of the paper. In (*b*) the view is rotated by about 90° so that the CD—OE1—OE2 plane is parallel to the plane of the paper.

CB—CG—CD (130.7°) and CG—CD—OE1 (139.2°) bond angles are larger than the ideal values by 18 and 21°, respectively, whereas the OE1—CD—OE2 (94.8°) angle is smaller by 29°. After the refinements, the strains on these angles are relieved in both the QM/MLF and *CNS*/MLF structures. However, the *CNS*/MLF structure shows a relatively large difference from the QM/MLF structure. Again, we computed the simluted-annealing omit map to aid our analysis, which shows that both the crystal and the QM/MLF structures fit reasonably well into the omit map, while this is not the case for the *CNS*/MLF structure, which is consistent with the suggestions from the local *R* values in Fig. 3.

Not only can QM refinement restrain structures where the electron density is insufficient to fully determine the atomic coordinates, but it also can reduce the deviations from ideality where the density is well resolved. In Fig. 6 we plot the structures of the main-chain atoms of Leu6 and the corresponding electron-density maps. The map calculated for the QM/MLF structure is very similar to that for the crystal structure and according to Fig. 3 the real-space *R* value and correlation coefficient are about the same for all three structures. However, the CA—C—O bond angle in the crystal structure is 111.7°, smaller than the ideal value of 120.8°. After restrained refinement, this deviation is corrected in the QM/MLF and *CNS*/MLF structures. It is clear from Fig. 6 that all three structures can fit snugly into the electron-density map; however, the QM/MLF and *CNS*/MLF structures would be energetically more plausible than the crystal structure. Another example of such adjustment to the main-chain atoms is the N—CA—C angle of Gly37, for which the real-space *R* value and correlation coefficient of the crystal structure are better than those of the QM/MLF structure, which are in turn better than those of the *CNS*/MLF structure. However, as the structures and maps shown in Fig. 7 indicate, the N—CA—C angle in the crystal structure is about 9° larger than the ideal value of 112.5° and this deviation is corrected in the QM/MLF and *CNS*/MLF structures. Again, the local *R* values show that the density of the QM/MLF structure for this residue fits the observed data better than that of the *CNS*/MLF structure.

4. Conclusions

In this paper, we have developed a method that uses energy restraints derived from quantum-mechanical calculations to refine crystal structures of proteins and have tested it on the crystal structure of the BPTI molecule.[1] In this method, the typical energy restraints derived from molecular mechanics have been replaced by those based on linear-scaling MO calculations. We adopted a protocol that determines the optimal weight at which the X-ray diffraction data are factored into the refinement process to maintain the compatibility of the refined structure with the X-ray signals and found that the weights determined by this criterion are slightly smaller than those that give the optimal free *R* values in the *CNS* refinements. The QM refinements gave comparable structures compared with the MM refinements in terms of the free *R* values and the real-space *R* values and correlation coefficients of the refined structures. It was also found that with the same restraints the MLF target function generally performed better than the residual one. The energetically restrained refinement was able to correct the deviations in most of the bond angles observed in the crystal structure from the ideal values derived from a statistical analysis of high-resolution structures of proteins and small molecules. Moreover, it was discovered that the semiempirical AM1 method led to systematic differences in certain bond lengths compared with the statistical data, which warrant future research in this direction.

[1] Supplementary data relating to this paper have been deposited with the IUCr electronic archive (Reference: GX5035). Details for accessing these data are given at the back of the journal.

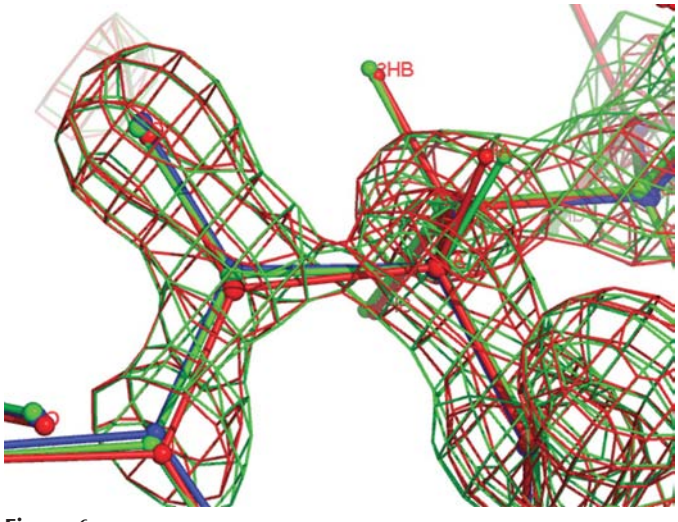

Figure 6
The main-chain atoms of Leu6 of the crystal structure (red), the QM/MLF structure (green) and the *CNS*/MLF structure (blue), together with the σ_A-weighted $2F_o - F_c$ electron-density maps based on the crystal structure (red) and the QM/MLF structure (green) contoured at the 2.4σ level.

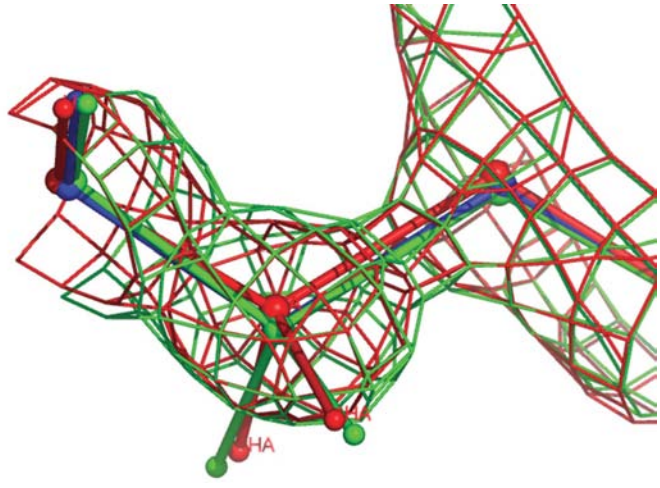

Figure 7
The main-chain atoms of Gly37 of the crystal structure (red), the QM/MLF structure (green) and the *CNS*/MLF structure (blue), together with the σ_A-weighted $2F_o - F_c$ electron-density maps based on the crystal structure (red) and the QM/MLF structure (green) contoured at the 2.4σ level.

It might be argued that under experimental conditions, especially in room-temperature crystals, X-ray diffraction data are the thermal averages over quite an extensive region of the conformational space, which is why an ensemble of structures obtained from the time-averaging technique is found to fit the X-ray data better than a single structure (Gros *et al.*, 1990; Schiffer & van Gunsteren, 1999). Knowledge-based parameters such as those constructed by Engh and Huber have the advantage that they can take entropy effects into account to some degree, whereas our approach is solely energy-driven. It is also worthwhile to note that in this study we have not attempted to treat solvent molecules and static disorder, which are traditionally difficult problems for any theoretical studies, but have only focused our attention on the protein molecule in the most probable configuration. However, as we have demonstrated, the advantages of QM refinement outweigh its current limitations and thus encourage future work that will make this method more practical for routine crystal structure determination. Furthermore, as computer power continues to increase, one might also imagine combining this method with advanced sampling and optimization techniques to explore the conformational space more extensively, which would better automate the refinement of protein crystal structures and prove valuable in improving the accuracy of X-ray crystallographic structure refinements.

This research was generously supported by the NSF (MCB-0211639) and the NIH (GM44974).

References

Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1997). *Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–474.

Brunger, A. T. & Adams, P. D. (2002). *Acc. Chem. Res.* **35**, 404–412.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Brünger, A. T., Karplus, M. & Petsko, G. A. (1989). *Acta Cryst.* A**45**, 50–61.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Catti, M. & Pavese, A. (1996). *Acta Cryst.* A**52**, 413–418.

Chandler, G. S., Figgis, B. N. & Li, Z. C. (2000). *Phys. Chem. Chem. Phys.* **2**, 3743–3751.

Chandler, G. S., Figgis, B. N., Reynolds, P. A. & Wolff, S. K. (1994). *Chem. Phys. Lett.* **225**, 421–426.

Deisenhofer, J. & Steigemann, W. (1975). *Acta Cryst.* B**31**, 238–250.

Dewar, M. J. S. (1969). *The Molecular Orbital Theory of Organic Chemistry*. New York: McGraw–Hill.

Dixon, S. L. & Merz, K. M. (1996). *J. Chem. Phys.* **104**, 6643–6649.

Dixon, S. L. & Merz, K. M. (1997). *J. Chem. Phys.* **107**, 879–893.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Field, M. J., Bash, P. A. & Karplus, M. (1988). *Biophys. J.* **53**, A46.

Gros, P., van Gunsteren, W. F. & Hol, W. G. J. (1990). *Science*, **249**, 1149–1152.

Hehre, W. J., Radom, L., Schleyer, P. von R. & Pople, J. A. (1986). *Ab Initio Molecular Orbital Theory*. New York: Wiley.

Hendrickson, W. A. (1985). *Methods Enzymol.* **115**, 252–270.

Housset, D., Benabicha, F., Pichon-Pesme, V., Jelsch, C., Maierhofer, A., David, S., Fontecilla-Camps, J. C. & Lecomte, C. (2000). *Acta Cryst.* D**56**, 151–160.

Huber, R., Kukla, D., Ruhlmann, A., Epp, O. & Formanek, H. (1970). *Naturwissenschaften*, **57**, 389.

Jack, A. & Levitt, M. (1978). *Acta Cryst.* A**34**, 931–935.

Jelsch, C., Teeter, M. M., Lamzin, V., Pichon-Pesme, V., Blessing, R. H. & Lecomte, C. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 3171–3176.

Krijn, M. P. C. M. & Feil, D. (1988). *J. Chem. Phys.* **89**, 4199–4208.

Krijn, M. P. C. M., Graafsma, H. & Feil, D (1988). *Acta Cryst.* B**44**, 609–616.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Lee, T. S., York, D. M. & Yang, W. T. (1996). *J. Chem. Phys.* **105**, 2744–2750.

Levine, I. R. (2001). *Quantum Chemistry*. Upper Saddle River, NJ, USA: Prentice–Hall.

Liu, H. Y., Elstner, M., Kaxiras, E., Frauenheim, T., Hermans, J. & Yang, W. T. (2001). *Proteins Struct. Funct. Genet.* **44**, 484–489.

Moulinier, L., Case, D. A. & Simonson, T. (2003). *Acta Cryst.* D**59**, 2094–2103.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.

Ryde, U. & Nilsson, K. (2003*a*). *J. Mol. Struct. (Theochem)*, **632**, 259–275.

Ryde, U. & Nilsson, K. (2003*b*). *J. Am. Chem. Soc.* **125**, 14232–14233.

Ryde, U. & Nilsson, K. (2003*c*). *J. Inorg. Biochem.* **96**, 39.

Ryde, U., Olsen, L. & Nilsson, K. (2002). *J. Comput. Chem.* **23**, 1058–1070.

Schiffer, C. A. & van Gunsteren, W. F. (1999). *Proteins Struct. Funct. Genet.* **36**, 501–511.

Swaminathan, S., Craven, B. M., Spackman, M. A. & Stewart, R. F. (1984). *Acta Cryst.* B**40**, 398–404.

Szabo, A. & Ostlund, N. S. (1996). *Modern Quantum Chemistry*. New York: McGraw–Hill.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–491.

Vaart, A. van der, Suarez, D. & Merz, K. M. (2000). *J. Chem. Phys.* **113**, 10512–10523.

Van Alsenoy, C., Yu, C. H., Peeters, A., Martin, J. M. L. & Schafer, L. (1998). *J. Phys. Chem. A*, **102**, 2246–2251.

Warshel, A. & Levitt, M. (1976). *J. Mol. Biol.* **103**, 227–249.

Wlodawer, A., Walter, J., Huber, R. & Sjolin, L. (1984). *J. Mol. Biol.* **180**, 301–329.

Wuthrich, K., Wider, G., Wagner, G. & Braun, W. (1982). *J. Mol. Biol.* **155**, 311–319.

Yang, W. (1991). *Phys. Rev. Lett.* **66**, 1438–1441.

Yang, W. T. & Lee, T. S. (1995). *J. Chem. Phys.* **103**, 5674–5678.

Yu, N. & Merz, K. M. (2004). *Mol. Phys.* **102**, 2545–2557.